Yan Xiong
Changchun Institute of Optics,
Fine Mechanics and Physics,
Chinese Academy of Sciences,
Space Robot Engineering Center,
Changchun, Jilin 130033, China;
University of Chinese Academy of Sciences,
Beijing 100049, China
e-mail: xiongyan16@mails.ucas.ac.cn

Liang Guo[1]
Professor
Changchun Institute of Optics,
Fine Mechanics and Physics,
Chinese Academy of Sciences,
Space Robot Engineering Center,
Changchun, Jilin 130033, China
e-mail: guoliang@ciomp.ac.cn

Defu Tian
Changchun Institute of Optics,
Fine Mechanics and Physics,
Chinese Academy of Sciences,
Space Robot Engineering Center,
Changchun, Jilin 130033, China;
University of Chinese Academy of Sciences,
Beijing 100049, China
e-mail: tiandefu19@mails.ucas.ac.cn

Application of Deep Reinforcement Learning to Thermal Control of Space Telescope

With the development of deep space exploration technology, thermal control systems for space telescopes are becoming increasingly complex, leading to the key parameters of conventional thermal control systems are difficult to adjust online automatically. To achieve these adjustments, this paper provided detailed verification of the application of deep reinforcement learning to space telescope thermal control from three perspectives: thermophysical modeling, intelligent sensing-based radiator, and online self-tuning of thermal control parameters. This paper presents a high-speed and high-precision thermophysical modeling strategy in MATLAB/SIMULINK with better computational efficiency than conventional approaches. And an intelligent sensing-based radiator is proposed that can realize autonomous regulation of the radiating cold plate by sensing the external space environment and the thermal load inside the spacecraft. A strategy for online self-tuning of the thermal control parameters based on deep reinforcement learning is also proposed. Theoretical and experimental results show that deep reinforcement learning thermal control (DRLPID) can achieve temperature control accuracy of 0.05 °C. The steady-state errors in the simulations were reduced by 22.7%, 37.4%, and 47.4% when compared with the reinforcement learning proportional–integral–derivative (PID), the neural network PID, and the fuzzy PID, respectively. The experimental steady-state errors were reduced by 20.4%, 32.5%, and 42.7%, respectively. [DOI: 10.1115/1.4051072]

Keywords: aerospace heat transfer, heat exchangers, thermal systems

1 Introduction

As an important type of space-based remote sensor, space telescopes have been used widely in military reconnaissance, resource exploration, disaster prediction, and other applications. With the rapid development of space remote sensing technology and the continuous improvements in space detection accuracy, the resolution requirements for space telescopes have become increasingly demanding [1–6]. On the one hand, an increase in the resolution means that the optical system must be larger in diameter and must offer greater accuracy; on the other hand, it also places more stringent requirements on the thermal control system in terms of its accuracy, robustness, and adaptability to the working environment [7]. In addition, as the complexity of the space missions of space telescopes increases, it becomes increasingly difficult for ground-based measurement and control systems to meet these demands. Factors such as the instability of the thermal environment in space and the possible failure of the space telescope itself also place further requirements for both autonomy and adaptability on the thermal control system for the space telescope. Therefore, it becomes necessary to study a new type of intelligent thermal control system.

At present, increasing numbers of researchers are beginning to conduct in-depth research into intelligent thermal control systems, mainly at the component and system levels, to improve the autonomous thermal control capabilities of spacecraft. Examples include NASA, who launched the Deep Space 1 (DS-1) probe as part of the New Millennium Project to complete verification of their autonomous control system during the flight control test [8–11]. Jia et al. [12] proposed an agent-based hierarchical hybrid structure for the autonomous thermal control method, leading to a thermal control system with pre-activity and adaptive and autonomous planning capabilities, to realize a spacecraft thermal control system for autonomous control of its primary technical purpose, and proposing an intelligent radiator, but this method is still at the theoretical stage and requires further in-depth research for engineering applications. Li et al. [13] described the working principle and fuzzy control algorithm of a new intelligent equivalent physical simulator which consists of a thermoelectric cooler, a plate-fin heat sink, a forced cooling fan, and an integrated fuzzy controller. This method applies fuzzy control to traditional proportional–integral–derivative (PID) controllers for space radiators, but lacks innovative designs in radiator architecture. In addition, Xin [14] introduced feedforward control using the "system thermal load" as a signal based on the existing PID feedback control system, which effectively reduced the dynamic deviation and overshoot of the system temperature characteristics and optimized the system's control performance, but the control efficiency of this method was rather slow and could not meet the practical engineering requirements. Wang et al. [15] developed a PID parameter optimization method based on finite element analysis (FEA), which is much better than the traditional methods. However, this method requires the FEA model of PID thermal control, which is extremely difficult and time-consuming to construct. Song et al. [16] provided a novel and effective method for high-precision thermal control of electronic devices by combining positive temperature coefficient (PTC) material with PID control strategy. The method employs PTC material for PID temperature control, which can effectively reduce the amount of process overshoot. However, the applicability of this method is too poor, especially without a complete theoretical system for PID parameter tuning. Grassi and Tsakalis [17] developed a frequency-loop-based PID tuning method that can be directly extended to multivariate PID cases. However, the simplicity of the method relies mainly on the linearity (convexity) in the

[1]Corresponding author.

Manuscript received December 19, 2020; final manuscript received April 30, 2021; published online June 17, 2021. Assoc. Editor: Prabal Talukdar.

parameter structure of the PID, which may not be preserved in tuning problems with different parameterizations or goals. Recently, we proposed an intelligent reinforcement learning-based thermal control strategy for adaptive self-tuning of PID parameters (reinforcement learning PID or RLPID) [18], which provides the PID thermal controller with the flexibility to tune the PID parameters for stable and precision thermal control. However, the method uses a low-speed algorithm that is difficult to bring to convergence and also generates large steady-state errors when the control object is changed. Thus, we then proposed an intelligent thermal control algorithm based on deep reinforcement learning (DRLTC) [19], which can quickly and accurately tune the PID parameters according to the control object and provide online current compensation based on the thermal system. However, the convergence of DRLTC is not particularly stable, especially without detailed verification analysis for thermal control systems with multiple zones heated simultaneously and coupled.

To solve the problems described above and achieve intelligent thermal control of space telescopes, an intelligent thermal control strategy based on deep reinforcement learning is proposed for space telescopes in this paper. In addition, the application of deep reinforcement learning to the thermal control of space telescopes will be verified in detail from three perspectives: thermophysical modeling, intelligent radiators based on multi-sensor fusion technology, and online self-tuning of the thermal control parameters. First, this paper presents thermophysical modeling of space telescopes in MATLAB/SIMSCAPE based on the node network method with high-speed and high accuracy, which was proposed and validated in paper [19], and this method provides improved computational efficiency when compared with the traditional approach of modeling using UG/TMG software. Second, to increase the heat dissipation efficiency of a radiator under passive thermal control, this paper proposes an intelligent sensor-based radiator based on multi-sensor fusion technology that operates by sensing the external space environment and the internal thermal payload of the spacecraft. It then uses the advantages of deep reinforcement learning in intelligent reasoning and decision-making to ensure that the thermal control system, with minimal power consumption, can realize the independent adjustment of the radiating cold plate, the tilt angle, and other parameters, thus achieving the ideal heat dissipation effect. Third, an intelligent thermal control policy for space telescopes (deep reinforcement learning PID thermal control system parameter tuning strategy or DRLPID) based on the deep deterministic policy gradient (DDPG) method is proposed. DDPG, which represents an important branch of reinforcement learning, is a data-driven control method that learns the system's mathematical model, achieves optimal system control based on the input and output data from the system, and then controls the thermal control system error based on the construction of the reward function. Therefore, the control parameters of the thermal control system of a space telescope could be adjusted automatically using deep reinforcement learning algorithms. Finally, the results of this work show that the performance of the proposed thermal control strategy is preferable to that of RLPID, backpropagation PID (BPPID), and fuzzy PID-based thermal control (FuzzyPID) [20,21].

The remainder of this paper is organized as follows. In Sec. 2, a thermophysical model of a space telescope in MATLAB and SIMULINK is provided. In Sec. 3, the smart sense radiator and DRLPID processes are described in detail. The simulated and experimental results are described in Secs. 4 and 5, respectively. Finally, Sec. 6 presents the conclusions from this study.

2 Thermophysical Model of Space Telescope

It is necessary to analyze the thermophysical model of a space telescope, which is highly complex, before applying the DRLPID approach to thermal control of that space telescope. At present, the node network method is the most widely used method for thermophysical modeling of spacecraft [22–24]. Depending on the characteristics of the space telescope, it can be divided into a large number of finite element units, where each of these units is treated as an equilibrium body and is used as a node. The following heat balance equation is applied to each node:

$$Q_1 = Q_2 + Q_3 + Q_4 + Q_5 \tag{1}$$

$$Q_1 = m_i c_i \frac{dT_i}{dt} \tag{2}$$

$$Q_2 = \sum_{j=1}^{N} (\alpha_{si} S \phi_{1i} + \alpha_{si} E_r \phi_{2i} \varepsilon_{li} E_e \varphi_{3i}) A_i \tag{3}$$

$$Q_3 = q_i \tag{4}$$

$$Q_4 = \sum_{j=1}^{N} D_{ji}(T_j - T_i) \tag{5}$$

$$Q_5 = \sum_{j=1}^{N} G_{ji}(T_j^4 - T_i^4) \tag{6}$$

where the subscripts i and j denote the node numbers; T denotes the temperature of the ith node; Q_1 is the value of internal energy variation of the node; m_i denotes the mass of the ith node; c_i denotes the specific heat of the ith node; t denotes the time; dT_i/dt is the temperature change rate of the ith node; Q_2 denotes the heating rate of the external heat flow absorbed by the ith node; α_{si} is the solar absorption coefficient of the surface of the ith node; S is the solar constant; E_r is the average reflection intensity of the earth to the solar radiation; ε_{li} is the emissivity of the surface of the ith node; E_e is the average infrared radiation intensity of the earth; ϕ_{1i}, ϕ_{2i}, and ϕ_{3i} are the angle factors of the surface of the ith node to the solar radiation, the earth albedo, and the infrared radiation of the earth, respectively; A_i is the area of the surface of the ith node; Q_3 and q_i denote the power of the internal heat source; Q_4 is the value of convective heat transfer between the node and the atmosphere environment; Q_5 is the value of radiation heat transfer between the node and the atmosphere environment; and D_{ji} and G_{ji} denote the linear heat conduction (i.e., the heat transfer coefficient) and the radiative heat conduction between node j and node i, respectively. In Eq. (1), the rate of change of the internal energy of the node is shown on the left side, and the right side shows the external heat flow absorbed by the node, the self-generated heat power, all linear thermal conduction-type heat transfer rates into the node, and all radiative-type heat transfer rates into the node, in that order.

The initial temperature conditions and thermal boundary conditions should be set as follows before calculating the node temperatures based on all the above equations

$$T(i)|_{t=0} = T_0(i) \tag{7}$$

$$T(i)|_{(S_c,S_\varepsilon)} = T_W(i) \tag{8}$$

where T_0 is the initial temperature; T_W is the temperature at a given boundary condition; S_c is the thermal boundary; and S_ε is the radiation boundary.

Figure 1 shows a schematic diagram and finite element model of the main mirror installation (MMI) in a space telescope system, and the relevant physical parameters are listed in Table 1. The main elements in the MMI include the main mirror, the intelligent electric heater, and the intelligent radiant cooling zone; the intelligent radiant cooling zone is used to cool the internal heat payload efficiently by arranging intelligent sensing emitters to adjust the cooling efficiency automatically according to the demands of the thermal control system. The MF501 NTC (provided by Chengdu Hongming Electronics Co., Ltd.) is a negative

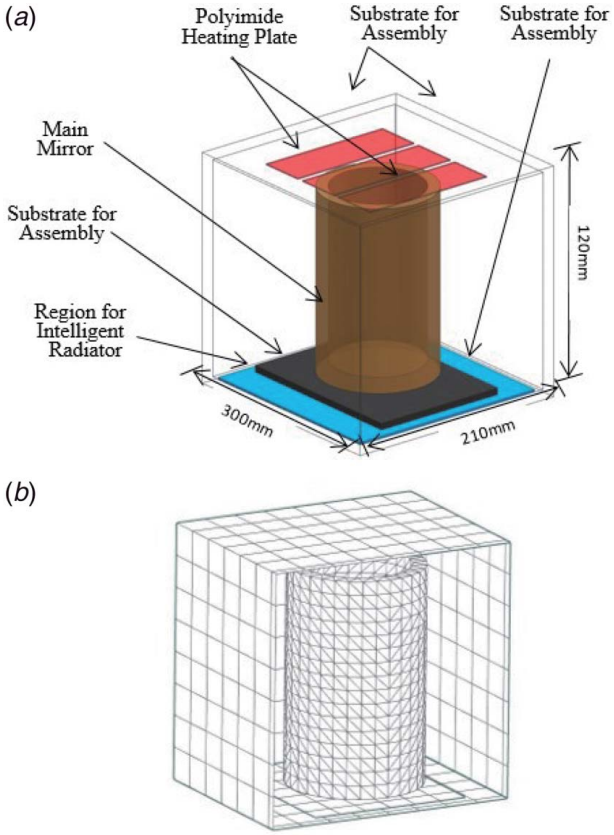

Fig. 1 Schematic diagram and finite element model of MMI

Table 1 Physical parameters of MMI

Parameter	Description	Value
ρ_{MMI}	Density	2637 kg/m^3
m_{MMI}	Quality	15.63 kg
k_{MMI}	Thermal conductivity	200 W/(m K)
c_{MMI}	Specific heat capacity	904 J/(kg K)
V_{MMI} (physical volume)	Volume	$0.3\times0.21\times0.12$ m^3

temperature coefficient thermistor that is used to perform high-precision temperature measurements with an accuracy of 5 mK [19]. Because the space telescopes system is in an ultra-low temperature state for a long period of time, the optical payload system is fully wrapped using multilayer insulation material to achieve a heat insulation effect.

As illustrated in Fig. 2(*a*), the node network method is used to divide the space telescope mounting box into six planes, which are named *L_Face*, *R_Face*, *U_Face*, *D_Face*, *F_Face*, and *B_Face*. The smart electric heater is arranged on the *U_Face*, while the main mirror and the smart radiant cooling zone are arranged on the *D_Face*, and insulation is arranged on all six surfaces. Each of these planes is then divided into 24 cells, individually named as Cells, as shown in Fig. 2(*b*). In addition, JRP represents an electric heater and MLI stands for the multilayer insulation. Based on the node network method, unit T41 was divided further into four units using SIMSCAPE [25,26] in MATLAB and SIMULINK and a thin plate unit body was created, as shown in Fig. 2(*c*). The method has been validated by several theoretical and experimental validations in Paper 1 [18] and Paper 2 [19] showing that the error between the thermophysical model built with SIMSCAPE in MATLAB and SIMULINK and the finite element model built in UG/TMG is always within 5%.

3 Intelligent Control Strategy Based on Deep Reinforcement Learning

3.1 Application to Self-Tuning of Parameters for Thermal Controllers. Figure 3 shows a schematic diagram of a PID adaptive thermal controller based on deep reinforcement learning. Because the thermal control system for the space-based optical remote sensors can only calculate the control parameter based on sampled deviation values and cannot use a continuous PID control algorithm directly because it requires a discrete approach, this paper proposes the concept of combining a deep reinforcement learning-based algorithm with a discrete positional PID controller [18,19,27–30]

$$\begin{aligned} u(t) &= K(t) + I(t) \\ &= k_P(t)x_1(t) + k_I(t)x_2(t) + k_D(t)x_3(t) + I(t) \\ &= k_P(t)error(t) + k_I(t)\sum_{j=0}^{k} error(j)T \\ &\quad + k_D(t)\frac{error(t) - error(t-1)}{T} + I(t) \\ &= k_P(t)(error(t) + \frac{T}{T_I}\sum_{j=0}^{k} error(j)) \\ &\quad + \frac{T_D}{T}(error(t) - error(t-1)) + I(t) \end{aligned} \tag{9}$$

where

$$k_I = \frac{k_P}{T_I} \tag{10}$$

$$k_D = k_P T_D \tag{11}$$

$$\begin{aligned} \int_0^t error(t)dt &\approx T\sum_{j=0}^{k} error(jT) \\ &= T\sum_{j=0}^{k} error(j) \end{aligned} \tag{12}$$

$$\begin{aligned} \frac{derror}{dt} &= \frac{error(kT)error((k-1)T)}{T} \\ &= \frac{error(t) - error(t-1)}{T} \end{aligned} \tag{13}$$

$$\begin{aligned} I(t) &= R(f) \\ &= R\left(\sqrt{\frac{2t^2}{\sum_{t=0}^{t} error(t) - error(t-1)^2}}\right) \end{aligned} \tag{14}$$

where T is the sampling time period; k is the sampling number, where $k=0, 1, 2, 3,\ldots$; $error(k-1)$ and $error(k)$ are the control errors of the thermal control system at the times $k-1$ and k, respectively; $K(t)=[K_P(t), K_I(t), K_D(t)]$ is the corresponding control parameter of the PID thermal controller at time t; $I(t)$ is the adaptive compensation of the PID thermal controller output current at time based on the deep reinforcement learning algorithm; and $R(f)$ is the current compensation function of the PID thermal controller based on the control error value, which is fitted approximately using the deep reinforcement learning algorithm.

As shown in Fig. 3, $y(t)$ and $y_d(t)$ represent the actual and expected system outputs, respectively. The system error $e(t)=y_d(t)-y(t)$ is converted into a system state vector $x(t)$ using a state converter known as a random action modifier (SAM). The Actor is used to perform strategic estimation and mapping of the system state variables to the recommended PID thermal controller parameters $K'(t)$ and the compensation current $I'(t)$. The output

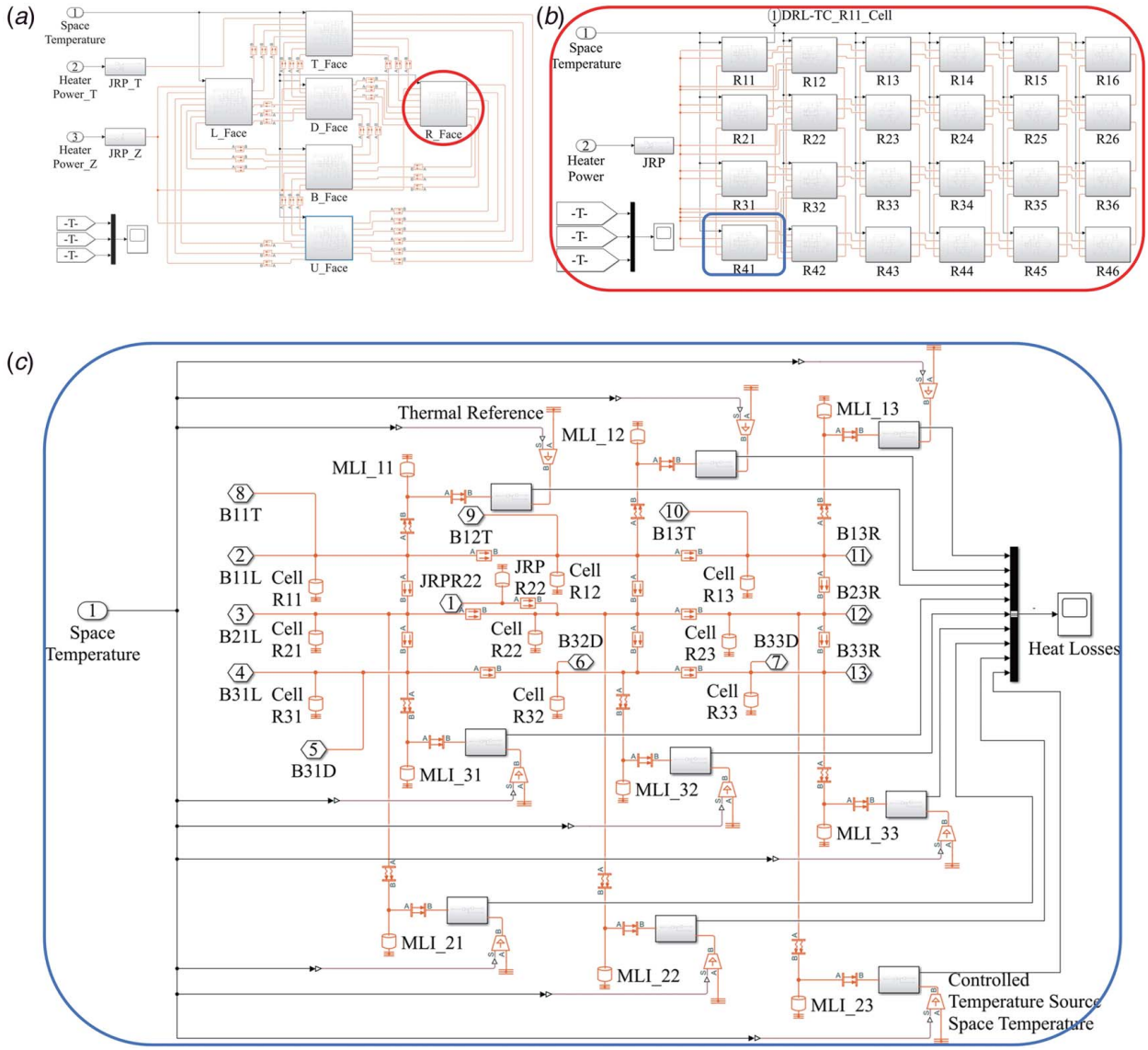

Fig. 2 Thermophysical model of the MMI

parameters of the Actor are not involved directly in the DDPG implementation of the PID thermal controller, but are corrected by the SAM based on value function estimation information provided by the Actor to obtain the actual PID thermal controller parameters $K(t)$ and the compensation current $I(t)$. The cumulative return value is denoted by $R(t)$. In addition, the decision effect of the DDPG was evaluated for each time period and the temporal difference (TD) error (internal reinforcement signal) δ_{TD} and an estimation function $V(t)$ were generated, where high values of δ_{TD} were fed directly to the Actor and the Critic and used to update their various parameters. Simultaneously, $V(t)$ was fed to SAM and used to correct the Actor's output.

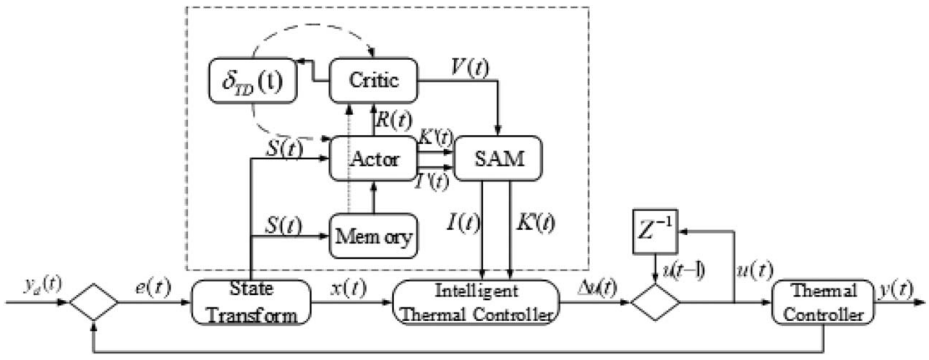

Fig. 3 Adaptive temperature controller based on deep reinforcement learning

The DDPG algorithm is a model-free, online, nonstrategic reinforcement learning method that was proposed by Lillicrap et al. [31]. This algorithm and the deep Q network (DQN) algorithm sample data in the same way and both algorithms use the experience replay sampling method; they sample randomly from their previous state transfer experience for training, thus solving the problem of the use of neural network representation value functions that are prone to algorithmic instability and other problems.

Because the DDPG algorithm is inspired by the idea of the DQN algorithm, it is necessary to describe the principle of the DQN algorithm in detail before the DDPG algorithm is introduced to enable better application of DDPG to online self-tuning of the thermal controller parameters. The DQN algorithm builds on the Q-learning algorithm by constructing two neural networks called the Actor Network and the Critic Network, and then creating two targets called the Target Actor Network and the Target Critic Network, which have the same network structures and are both slow to update. The algorithm uses the network to produce the target value, rather than use a Q-table to approximate the optimal Q-value, thus solving the problem that the Q-table cannot store state-action pairs in a high-dimensional continuous state, and exerts the deep learning ability to process the high-dimensional data. The DQN uses the Q function as the value evaluation function, and its renewal equation is given as follows:

$$Q(s, a) = Q(s_t, a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (15)$$

where $Q(s_t, a_t)$ denotes the current state Q-value, $\alpha \in [0, 1]$ denotes the learning rate of the DQN algorithm, $\gamma \in [0, 1]$ denotes the DQN algorithm discount factor, r_t denotes the payoff value, and $Q(s_{t+1}, a)$ denotes the line of Q data used for all actions in the next state.

The DQN uses a neural network as a Q-value network with the parameter ω and

$$Q(s, a, \omega) \approx Q^{\pi}(s, a) \quad (16)$$

The Critic Network uses the mean square error to define the minimum loss function as

$$L(\theta) = E[(Q_{\text{Target}} - Q^{\tau}(s_t, a; \theta))^2] \quad (17)$$

where q represents the neural network parameters and the objective function is

$$Q_{\text{Target}} = r + \gamma \max_a Q(s_{t+1}, a; \theta) \quad (18)$$

The Q-value network parameters ω are then computed for the gradient of the loss function as follows:

$$\frac{\partial L(\omega)}{\partial \omega} = E\left[(r + \gamma \bullet \max_a Q(s_{t+1}, a, \omega)) \frac{\partial Q(s_t, a, \omega)}{\partial \omega}\right] \quad (19)$$

where the value of $\partial Q(s_t, a, \omega)/\partial \omega$ is calculated by the neural network; the network parameters can be updated using the stochastic gradient descent (SGD) method, and the optimal Q-value is finally obtained.

Figure 4 shows a diagram of the self-tuning of the parameters based on the DDPG approach. Similar to the DQN algorithm, the DDPG algorithm also uses an Actor-Critic algorithm framework, which shows a flowchart of the Actor-Critic reinforcement learning structure. The general operation of the framework is described as follows.

The Actor uses the policy gradient for policy learning to select the thermal control strategies in the current given environment, while the Critic uses policy evaluation to evaluate the value function to generate signals for use in the evaluation of the Actor's actions. When the thermal control strategy is being planned, the external space environment data obtained from the spacecraft's thermal sensors and the internal heat load data are input into the Actor Network, which outputs the thermal control strategy to be adopted by the intelligent thermal control system;

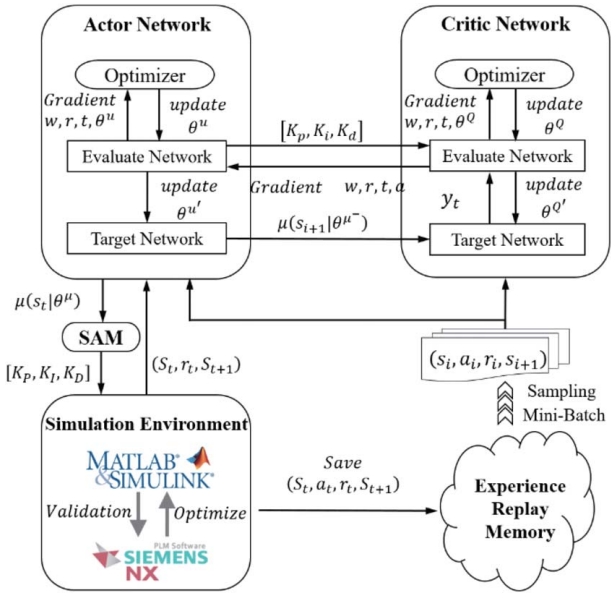

Fig. 4 Diagram of self-tuning of the parameters based on DDPG

the Critic network then inputs the internal and external thermal environment states of the spacecraft and the proposed thermal control strategy, and subsequently outputs the corresponding Q-value required for the evaluation.

In the DDPG algorithm, both the Actor and the Critic are represented by a deep neural network (DNN) and the Actor Network and the Critic Network perform function approximations on deterministic policy θ^{μ} and the value-action function Q with the parameters θ^{μ} and θ^{Q}, respectively. When the algorithm is updated iteratively, we initially accumulate the sample data in the experience pool until the minimum number of batches is reached; we then use the sample data to update the Critic Network and update the parameters θ^{Q} through the loss function to obtain the gradient for the relative action θ^{μ} of the objective function; finally, we use the Adam optimizer to update θ^{μ}.

It continues to be measured using a function J that is defined as follows when measuring a good or bad thermal control strategy

$$J_{\beta}(\mu) = \int_S \rho^{\beta}(s) Q^{\mu}(s, \mu(s)) ds = E_{s\sim\rho^{\rho}}[Q^{\mu}(s, \mu(s))] \quad (20)$$

where S denotes the environmental state of the spacecraft's thermal control system, ρ^{β} denotes the distribution function, and $Q^{\mu}(s, \mu(s))$ denotes the Q-value generated by the thermal control system in state S when it selects the thermal control strategy based on strategy μ. The algorithm is trained along the directions in which the loss function is maximized and minimized. The algorithm is then trained in the direction that maximizes $J_{\beta}(\mu)$ and minimizes the loss function. The DDPG algorithm uses the stochastic gradient ascent (SGA) random gradient to update the parameter θ^{μ} in the policy network μ. The Q network updates the parameter θ^{Q} in the same manner as the DQN algorithm.

3.2 Applications of Intelligent Sensing Radiator. The optimization of spacecraft thermal control systems based on both heat transfer and control science will make the task of thermal control of spacecrafts much more effective. Currently, many scholars have conducted a lot of research in the field of efficient cooling and heat transfer technologies for spacecraft and electronic devices, such as the microchannel heat exchangers developed by the Jet Propulsion Laboratory (JPL) and NASA, which can be integrated with the packaging structure of electronic devices [32,33],

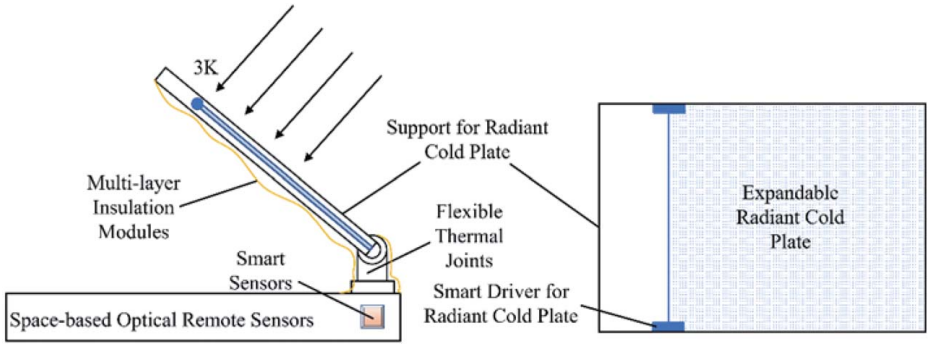

Fig. 5 Schematic diagram of an intelligent sensor-based radiator structure, which mainly consists of an unfolded radiating cold plate, a radiating cold plate intelligent driver, a radiating cold plate support frame, a flexible thermal joint, an intelligent perceptron, and a multilayer insulation assembly

single-phase and multi-phase flow circuits driven by micro-pumps [34,35], and micro louvers controlled by electrical signals [36,37], but the research on advanced active control technologies for various efficient cooling devices is still relatively lacking.

The thermal control system thus faces a huge challenge and to solve these problems, this paper proposes an intelligent autonomous regulation strategy based on deep reinforcement learning for radiators and a first sample structure diagram is designed that has some reference significance for future research and development of intelligent radiators.

Figure 5 shows a schematic diagram of an intelligent sensor-based radiator structure, which mainly consists of an unfolded radiating cold plate, a radiating cold plate intelligent driver, a radiating cold plate support frame, a flexible thermal joint, an intelligent perceptron, and a multilayer insulation assembly. By sensing the external thermal environment in space and the thermal load inside the space telescope and using intelligent decision-making algorithms, the intelligent sensor infers the heat dissipation area required for the radiation cooling plate and the mounting angle needed between the support frame and the remote sensor to ensure that the thermal control system can achieve stable control of the space telescope with the lowest possible power consumption. A flexible thermal joint is used to adjust the mounting angle of the radiant cooling panel support frame independently by receiving independent perception and reasoning decisions from the intelligent sensor. The unfolding radiant cooling panel then adjusts the radiant cooling panel's heat dissipation area in accordance with the independent adjustment of the intelligent driver of the radiant cooling panel and in sufficient time to adjust the heat dissipation and cooling efficiency of the thermal control system based on the reasoning decisions of the intelligent sensor to finally achieve the ideal heat dissipation effect.

4 Simulation

To verify whether the intelligent thermal control strategy based on deep reinforcement learning proposed in this paper can adjust the system autonomously and then reach the ideal state via unsupervised adjustment when the external thermal environment or the internal thermal load of a space telescope changes, the strategy is applied to temperature control of the MMI in an optical payload system. The physical parameters related to the MMI are listed in Table 1. There is strong thermal coupling between the MMI and the main mirror, which means that a temperature change in the MMI will affect the temperature stability of the main mirror directly; this represents a difficult challenge for the thermal control system because the temperature control accuracy of the main mirror is required to reach 0.1 °C. As shown in Fig. 1, multilayer insulation with an emissivity of 0.69 and an infrared absorbance of 0.02 was applied to the outer surface of the mounting box to reduce the effects of the external thermal environment on the internal temperature of the box.

The simulations performed in this study were divided into two heating phases to evaluate the performances of the DRLPID, RLPID, FuzzyPID, and BPPID approaches under the action of random disturbances from the external thermal environment. The total duration of the experiment is 3000 s. Over the period from 0 to 1500 s, the temperature in the thermally coupled zone is increased gradually from the initial temperature of 21 °C to 22 °C and this is used as the operating temperature for this phase, which is referred to as case A. From 1501 s to 3000 s, the temperature in the thermally coupled zone rises gradually to 23 °C, and this is used as the operating temperature at this stage; this stage is called case B. As a rule, two irregular 1–2 °C temperature disturbance signals are applied to these two cases at the installation position of the main mirror to verify the robustness of the proposed intelligent control algorithm. An active temperature control loop is located on the opposite side of the radiation dissipation area and is controlled by DRLPID. The parameters related to the simulation environment are listed in Table 2.

To verify the control effects of the intelligent thermal controller proposed in this paper, the simulation results of the RLPID designed by Xiong et al. [18], the FuzzyPID designed by Carvajal et al. [20], and the BPPID designed by Chen and Huang [21] are compared

Table 2 Simulation environment parameters

Parameter	Description	Value
P_{atm}	Atmospheric pressure	101.325 kPa
T_{inside}	The temperature inside the incubator	24.65 °C
$T_{outside}$	The temperature outside the incubator	25.55 ± 0.3 °C
P_{Heat}	Heating power	0–12.484 W
$\varepsilon_{\mathrm{MLI}}$	Emission coefficient of multilayer	0.69
a_{MLI}	Absorption coefficient of multilayer	0.02

Table 3 Parameters of DRLPID

Parameter	Description	Value
α_A	Learning rate of Actor	0.001
α_C	Learning rate of Critic	0.02
ε	Tolerant error band	0.001
γ	Discount factor	0.97
N_A, N_C	Batch size of Actor and Critic	256
M_A, M_C	Memory capacity of Actor and Critic	50,000
S_A	Update steps of Actor	15,000
S_C	Update steps of Critic	13,000

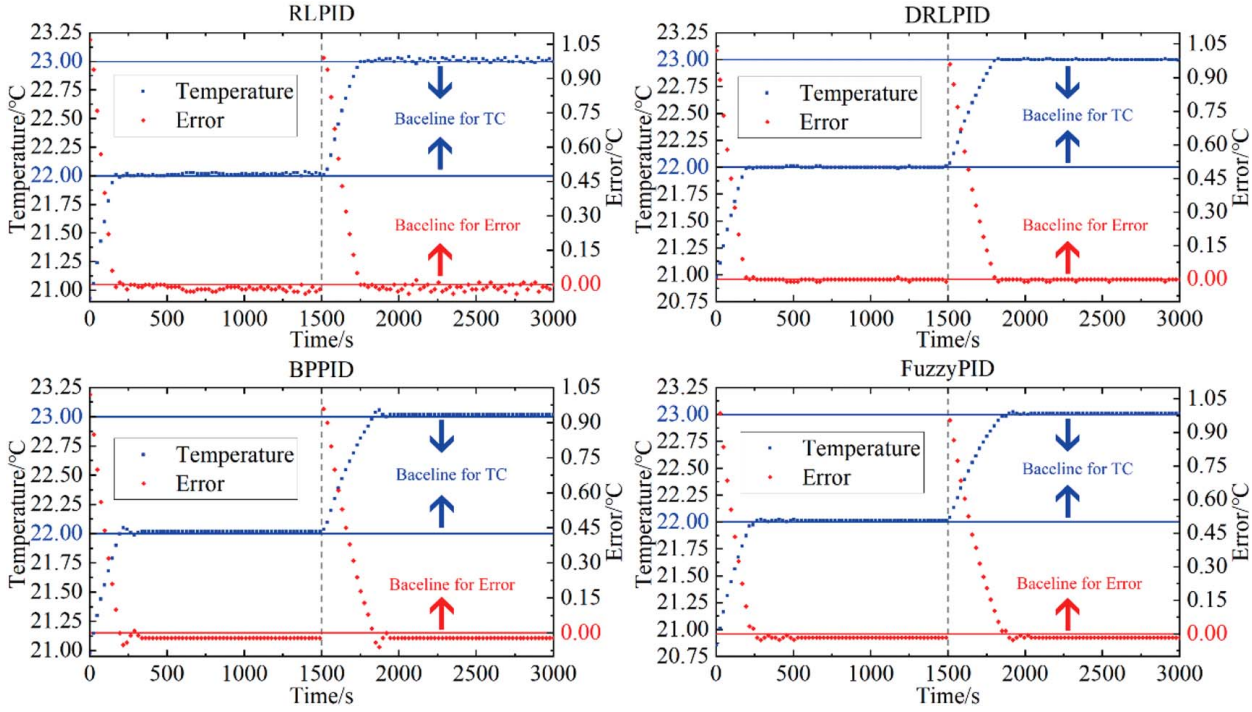

Fig. 6 Simulation results of MMI in the specified environment

and analyzed with respect to the simulation results obtained for DRLPID; the relevant parameters for DRLPID are presented in Table 3. The thermophysical model was built in SIMULINK using SIMSCAPE, and the control algorithm was designed using MATLAB. The simulation results are shown in Fig. 6.

As shown in Fig. 6, when the temperature increases gradually to reach the target temperature value, the RLPID thermal controller has an overshoot of up to 0.091 °C and its convergence effect is very poor, remaining constantly in the oscillation state despite a small fluctuation range. The FuzzyPID thermal controller has an overshoot of approximately 0.089 °C, and its static error exceeds 0.025 °C. For the BPPID thermal controller, the overshoot was 0.11 °C and the static error was 0.021 °C, and it cannot meet the thermal control indicators of the main mirror thermal control system. The static error of the DRLPID thermal controller proposed in this paper is only 0.052 °C; there is almost no overshoot in the temperature adjustment stage, and the temperature control stage shows less fluctuation when compared with the other controllers. When compared with the RLPID thermal controller, the BPPID thermal controller, and the FuzzyPID thermal controller, the temperature control accuracy of the DRLPID thermal controller proposed in this paper shows improvements of 22.7, 37.4, and 47.4%, respectively.

5 Experiments

For our experiments, we used the same MMI that was used for the simulations, and the relevant parameters of this MMI are given in Tables 1 and 2. An NTC thermistor called the MF501 NTC that can achieve temperature accuracy of 5 mK was used to perform the high-precision temperature measurements at critical

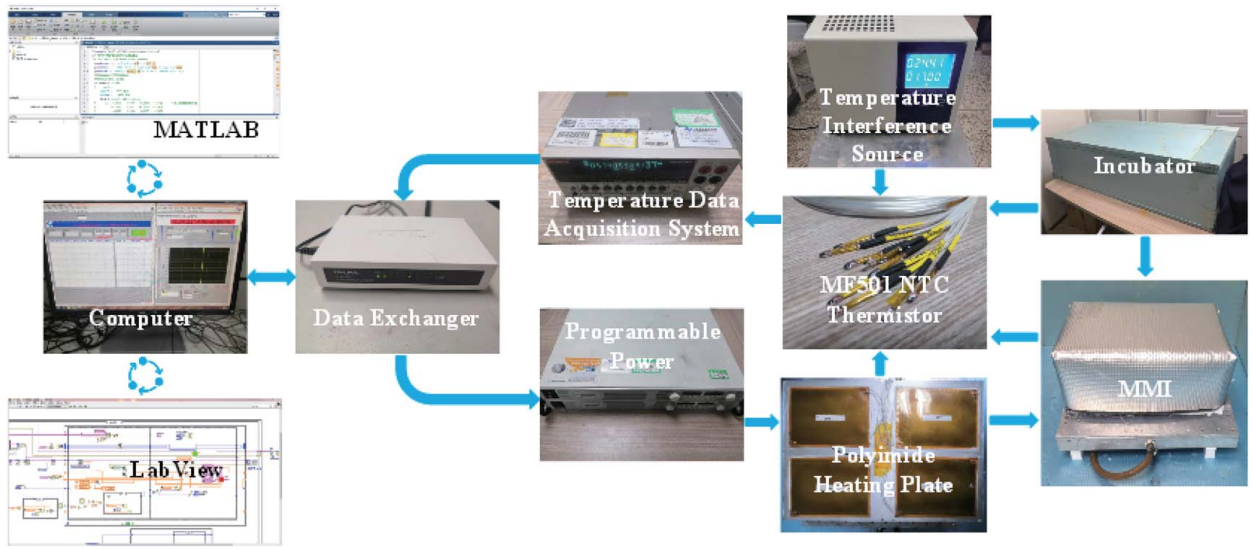

Fig. 7 Thermal equilibrium test devices and platforms for MMI, and experiment flow

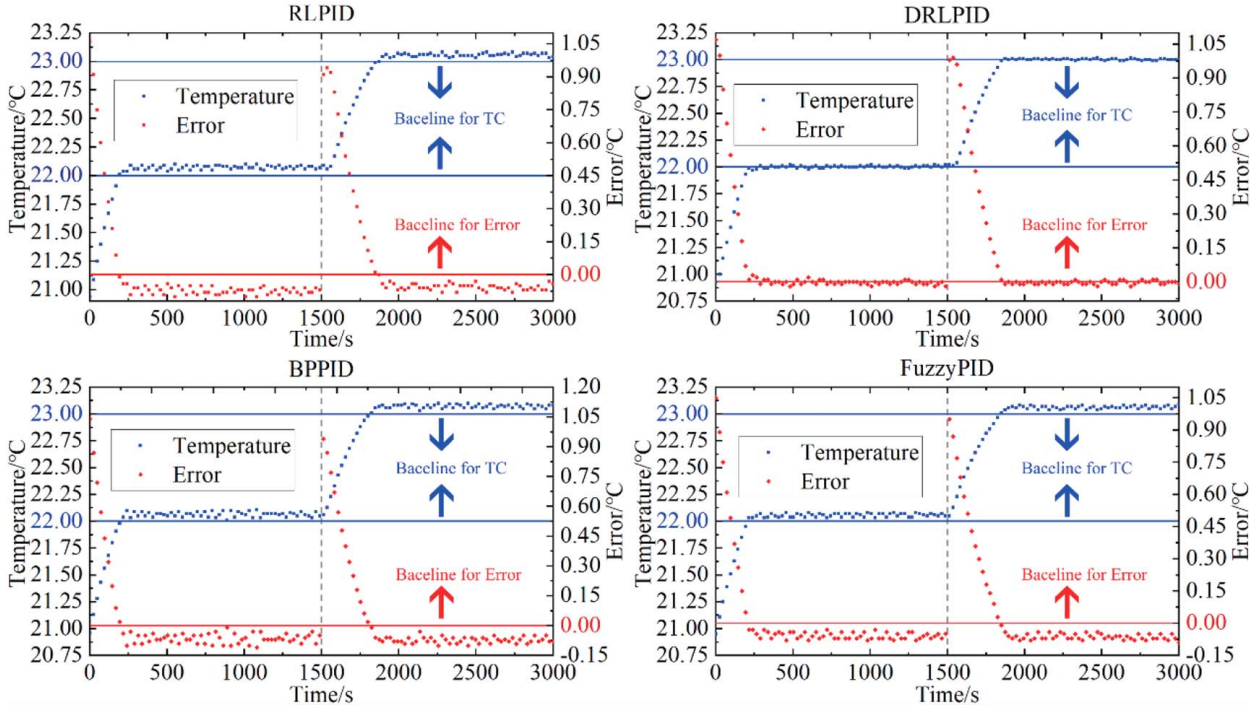

Fig. 8 Experimental results of MMI in the specified environment

locations. As shown in Fig. 7, a heating zone was placed above the MMI to prevent the high temperatures generated during heating of the polyimide heating element from affecting the imaging quality of the primary mirror. The MMI was placed in a holding tank to insulate it from the external environment, thus eliminating any effects from the external environment. To verify the robustness of the various control methods, a 1–2 °C random temperature interference signal was applied in the experiment at the same locations used in the simulations.

The algorithm was developed and designed in MATLAB, and LABVIEW was used to perform joint control [38,39]. The experimental flow is illustrated in Fig. 7.

The temperature data from the important nodes in the MMI were collected in real-time using the temperature data acquisition system and were then sent to the DRLPID algorithm system in MATLAB via a data switch. The DRLPID algorithm system then made intelligent decisions by reasoning and sent these decisions to the programmable power supply through the data switch to be used to control the power supply. As shown in Fig. 8, the BPPID thermal controller had an overshoot of 0.15 °C, a static error of 0.067 °C, a temperature amplitude of 0.045 °C, and found it difficult to reach convergence, always remaining in a fluctuating state. The RLPID thermal controller also showed large temperature fluctuations and had an overshoot of 0.11 °C. Although the maximum temperature fluctuation of the FuzzyPID thermal controller was only 0.096 °C, its static error reached 0.079 °C. The maximum temperature fluctuation of the DRLPID was only 0.075 °C, and its static error was only 0.057 °C. When compared with the RLPID, the BPPID, and the FuzzyPID, the temperature control accuracy was improved by 20.4, 32.5, and 42.7%, respectively, when using the DRLPID.

6 Conclusions

In this paper, a self-tuning strategy for thermal control parameters based on deep reinforcement learning and an intelligent autonomous regulation strategy for radiators based on deep reinforcement learning are proposed. The feasibility and the advantages of the application of deep reinforcement learning to the thermal control of space telescopes are verified in detail from three perspectives: thermal physics modeling, intelligent perception of the emitters, and online self-tuning of the thermal control parameters. An adaptive temperature controller based on deep reinforcement learning is used to perform simulation analyses and experimental verification of the thermal control system of the main mirror mounting box in an optical payload system, and the following conclusions can be drawn from the analysis of the research results:

(1) The node network method proposed in this paper for thermophysical modeling of space telescopes in SIMULINK using SIMSCAPE can integrate the control algorithms effectively, thus enabling simultaneous finite element analysis and control algorithm verification of thermophysical models of space telescopes and thereby improving the modeling efficiency and control algorithm development efficiency for their thermal control systems.
(2) The feasibility and prospects for the application of deep reinforcement learning algorithms in self-tuning of the thermal control system parameters of a space telescope have been demonstrated by simulation and testing of the thermal control system of an MMI in a payload system.
(3) The DRLPID proposed in this paper is applied to the space telescope's adaptive thermal controller. The simulation temperature control accuracy can reach 0.02 °C and the steady-state error of the experiment is only 0.05 °C, which means that the DRLPID offers the advantages of higher accuracy, increased robustness, and a smaller steady-state error when compared with the various traditional adaptive PID thermal controllers.
(4) The strategy used for self-tuning of the thermal control parameters based on deep reinforcement learning and the intelligent autonomous regulation strategy for the radiators based on deep reinforcement learning proposed in this paper can realize the intelligent autonomous regulation of the thermal control system under no supervision, which greatly improves the efficiency of the thermal control system and can provide novel ideas for the intelligent and autonomous development of the thermal control system of future space telescopes.

Acknowledgment

This work received funding from the National Natural Science Foundation of China (Grant No. 61605203) and the Youth Innovation Promotion Association of Chinese Academy of Sciences (Grant No. 2015173).

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

Nomenclature

t = time
a_{MLI} = absorption coefficient of multilayer
a_t = the action at moment t
c_i = specific heat capacity of the ith node
k_{MMI} = thermal conductivity of MMI
m_i = mass of the i node
m_{MMI} = quality of MMI
q_i = power of the internal heat source of the ith node
r_t = payoff value
s_t = system state at moment t
y_t = actual system outputs
A_i = the area of the surface of the ith node
D_{ji} = linear heat conduction (i.e., the heat transfer coefficient) between node j and node i
E_e = the average infrared radiation intensity of the earth
E_r = the average reflection intensity of the earth to the solar radiation
G_{ji} = the radiative heat conduction between node j and node i
M_A = memory capacity of the Actor agent
M_C = memory capacity of the Critic agent
N_A = batch size of the Actor agent
N_C = batch size of the Critic agent
P_{atm} = atmospheric pressure
P_{heat} = heating power
Q_1 = the value of internal energy variation of the ith node
Q_2 = heating rate of the external heat flow absorbed by the ith node
Q_3 = power of the internal heat source
Q_4 = the value of convective heat transfer between the node and the atmosphere environment
Q_5 = the value of radiation heat transfer between the node and the atmosphere environment
S_A = update steps of Actor
S_c = thermal boundary
S_C = update steps of Critic
S_ε = radiation boundary
T_i = temperature of the ith node
T_j = temperature of the jth node
T_{inside} = the temperature inside the incubator
$T_{outside}$ = the temperature outside the incubator
$DRLPID$ = intelligent thermal control policy for space telescopes based on the DDPG
$error_t$ = control errors of the thermal control system at the time t
$e(t)$ = system error
$I(t)$ = adaptive compensation of the PID thermal controller output current at time t based on the DRL
$I'(t)$ = compensation current
$K(t)$ = control parameters of the PID thermal controller at time t
$K'(t)$ = recommended PID thermal controller parameters from the Actor agent
$R(f)$ = current compensation function of the PID thermal controller based on the control error value
$R(t)$ = cumulative return value
$T_0(i)$ = initial temperature of the ith node
$T_w(i)$ = temperature at a given boundary condition
$u(t)$ = control strategy
V_{MMI} = volume of MMI
$V(t)$ = estimation function
$y_d(t)$ = expected system outputs
α_A = learning rate of the Actor
α_C = learning rate of the Critic
α_{su} = solar absorption coefficient of the surface of the ith node
γ = discount factor
δ_{TD} = temporal difference
ε = tolerant error band
$\varepsilon_{\mathrm{MLI}}$ = emissivity coefficient of multilayer
θ^μ = deterministic policy
θ^Q = value-action function
ρ^β = distribution function

References

[1] Lei, Z., Ming, L., Danni, L., Zhao, Z., and Liujun, Z., 2017, "Thermal Optics Property Study and Athermal Design on Optical Window of IR Aiming Device Reliability Testing System," Optik, **136**, pp. 586–594.
[2] Kumar Rai, P., Rao Chikkala, S., Adoni, A. A., and Kumar, D., 2015, "Space Radiator Optimization for Single-Phase Mechanical Pumped Fluid Loop," ASME J. Therm. Sci. Eng. Appl., **7**(4), p. 041021.
[3] Phoenix, A. A., and Wilson, E., 2018, "Adaptive Thermal Conductivity Metamaterials: Enabling Active and Passive Thermal Control," ASME J. Therm. Sci. Eng. Appl., **10**(5), p. 051020.
[4] Li, S., Chen, L., and Yang, Y., 2020, "Thermal Design and Test Verification of the Solar X-Ray and Extreme Ultraviolet Imager," Optik, **203**(2018), p. 164017.
[5] Huang, P. G., and Doman, D. B., 2018, "Thermal Management of Single- and Dual-Tank Fuel-Flow Topologies Using an Optimal Control Strategy," ASME J. Therm. Sci. Eng. Appl., **10**(4), p. 041019.
[6] Gao, Y., Zhang, B., Chen, L., Xu, B., and Gu, G., 2019, "Thermal Design and Analysis of the High Resolution MWIR/LWIR Aerial Camera," Optik, **179**, pp. 37–46.
[7] Jihui, L., Shuangli, H., Jiaqi, W., L, E., and Jun, W., 1999, "Thermal Analysis and Thermal Control Techniques of Space Camera," Opt. Precis. Eng., **7**(6), pp. 36–41.
[8] Choi, M. K., 2004, "Method of Generating Transient Equivalent Sink and Test Target Temperatures for Swift BAT," Collection of Technical Papers—2nd International Energy Conversion Engineering Conference, Vol. 3, pp. 1377–1384.
[9] John Anger Richmond, L., Colonel John Keesee, U., and Retired, U., 2010, "Adaptive Thermal Modeling Architecture for Small Satellite Applications," Doctoral dissertation, Massachusetts Institute of Technology.
[10] Galski, R. L., De Sousa, F. L., Ramos, F. M., and Muraoka, I., 2007, "Spacecraft Thermal Design With the Generalized Extremal Optimization Algorithm," Inverse Problems Sci. Eng., **15**(1), pp. 61–75.
[11] Lemmen, M., Kouwen, J., Koorevaar, F., and Pennings, N., 2004, "In-Flight Results of the Sciamachy Optical Assembly Active Thermal Control System," SAE Technical Paper No. 2004-01-2357.
[12] Jia, L., Yunze, L., Jiaxun, Z., Peiguang, W., and Jun, W., 2009, "Development of the Spacecraft MEMS Autonomous Thermal Control System Based on Intelligent Agent," Spacecraft Environ. Eng., **26**(6), pp. 574–579.
[13] Li, Y.-Z., Lee, K.-M., and Wang, J., 2009, "Intelligent Equivalent Physical Simulator for Nanosatellite Space Radiator," 2009 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, IEEE, Singapore, July 14–17, IEEE, pp. 504–509.
[14] Xin, N. X. Z. J. Z., 2008, "Feed Forward PID Control of Satellite Single-Phase Fluid Loop Thermal Control System," Chin. Space Sci. Technol., **28**(4), p. 4.
[15] Wang, Z., Yin, Z., and Xiong, Y., 2010, "Temperature Control and PID Parameters Optimization Based on Finite Element Model," 2010 International Conference on Electrical and Control Engineering, IEEE, Wuhan, China, June 25–27, pp. 2241–2244.
[16] Song, J., Cheng, W., Xu, Z., Yuan, S., and Liu, M., 2016, "Study on PID Temperature Control Performance of a Novel PTC Material With Room Temperature Curie Point," Int. J. Heat Mass Transfer, **95**, pp. 1038–1046.
[17] Grassi, E., and Tsakalis, K., 2000, "PID Controller Tuning by Frequency Loop-Shaping: Application to Diffusion Furnace Temperature Control," IEEE Trans. Control Syst. Technol., **8**(5), pp. 842–847.

[18] Xiong, Y., Guo, L., Huang, Y., and Chen, L., 2019, "Intelligent Thermal Control Strategy Based on Reinforcement Learning for Space Telescope," J. Thermophys. Heat Transfer, **34**(1), pp. 1–8.
[19] Xiong, Y., Guo, L., Wang, H., Huang, Y., and Liu, C., 2020, "Intelligent Thermal Control Algorithm Based on Deep Deterministic Policy Gradient for Spacecraft," J. Thermophys. Heat Transfer, **34**(4), pp. 1–13.
[20] Carvajal, J., Chen, G., and Ogmen, H., 2000, "Fuzzy PID Controller: Design, Performance Evaluation, and Stability Analysis," Information Sci., **123**(3), pp. 249–270.
[21] Chen, J., and Huang, T. C., 2004, "Applying Neural Networks to On-Line Updated PID Controllers for Nonlinear Process Control," J. Process Control, **14**(2), pp. 211–230.
[22] Richmond, J. A., 2010, "Adaptive Thermal Modeling Architecture for Small Satellite Applications," Ph.D. thesis, Massachusetts Institute of Technology.
[23] Lyon, R., Sellers, J., and Underwood, C., 2002, "Small Satellite Thermal Modeling and Design at USAFA: FalconSat-2 Applications," IEEE Aerospace Conf. Proc., **7**, pp. 3391–3399.
[24] Kov`acs, R., and J´ozsa, V., 2018, "Thermal Analysis of the SMOG-1 PocketQube Satellite," Appl. Therm. Eng., **139**, pp. 506–513.
[25] The Mathworks Inc., 2018, MATLAB—MathWorks.
[26] Ahmad, P., Ali Mohammadi, M. S., and Parvaresh, A., 2012, "A New Mathematical Dynamic Model for HVAC System Components Based on Matlab/Simulink," Int. J. Innovative Technol. Exploring Eng., **1**(2), pp. 1–6.
[27] Shipman, W. J., and Coetzee, L. C., 2019, "Reinforcement Learning and Deep Neural Networks for PI Controller Tuning," IFAC-PapersOnLine, **52**(14), pp. 111–116.
[28] Shang, X., Ji, T., Li, M., Wu, P., and Wu, Q., 2013, "Parameter Optimization of PID Controllers by Reinforcement Learning," 2013 5th Computer Science and Electronic Engineering Conference (CEEC), IEEE, Colchester, UK, Sept. 17–18, pp. 77–81.
[29] Lambert, N. O., Drew, D. S., Yaconelli, J., Levine, S., Calandra, R., and Pister, K. S., 2019, "Low-Level Control of a Quadrotor With Deep Model-Based Reinforcement Learning," IEEE Robot. Automation Lett., **4**(4), pp. 4224–4230.
[30] El Hakim, A., Hindersah, H., and Rijanto, E., 2013, "Application of Reinforcement Learning on Self-Tuning PID Controller for Soccer Robot Multi-Agent System," 2013 Joint International Conference on Rural Information & Communication Technology and Electric-Vehicle Technology (rICT & ICeV-T), IEEE, Bandung, Indonesia, Nov. 26–28, pp. 1–6.
[31] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D., 2016, "Continuous Control With Deep Reinforcement Learning," 4th International Conference on Learning Representations, ICLR 2016—Conference Track Proceedings, International Conference on Learning Representations, ICLR, San Juan, Puerto Rico, May 2–4.
[32] Lee, P.-S., Garimella, S. V., and Liu, D., 2005, "Investigation of Heat Transfer in Rectangular Microchannels," Int. J. Heat Mass Transfer, **48**(9), pp. 1688–1704.
[33] Westheimer, D., and Tuan, G., 2005, "Active Thermal Control System Considerations for the Next Generation of Human Rated Space Vehicles," 43rd AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, Jan. 10–13, p. 342.
[34] Paris, A. D., Birur, G. C., and Green, A. A., 2002, "Development of MEMS Microchannel Heat Sinks for Micro/Nano Spacecraft Thermal Control," ASME Int. Mech. Eng. Congress Expos., **36428**, pp. 25–31.
[35] Birur, G. C., Sur, T. W., Paris, A. D., Shakkottai, P., Green, A. A., and Haapanen, S. I., 2001, "Micro/Nano Spacecraft Thermal Control Using a MEMS-Based Pumped Liquid Cooling System," Microfluidics and BioMEMS, San Francisco, CA, Sept. 28, vol. 4560, pp. 196–206.
[36] Osiander, R., Champion, J., Darrin, M., Allen, J., Douglas, D., and Swanson, T., 2002, "Micro-Machined Shutter Arrays for Thermal Control Radiators on ST5," 40th AIAA Aerospace Sciences Meeting & Exhibit, Reno, NV, Jan. 14–17, p. 359.
[37] Osiander, R., Firebaugh, S. L., Champion, J. L., Farrar, D., and Darrin, M. G., 2004, "Microelectromechanical Devices for Satellite Thermal Control," IEEE Sens. J., **4**(4), pp. 525–531.
[38] National Instruments, 2017, LabVIEW—National Instruments.
[39] Yu, Y., Zhang, Y., Yuan, X., and Hou, Q., 2014, "A LabVIEW-Based Real-Time Measurement System for Polarization Detection and Calibration," Optik, **125**(10), pp. 2256–2260.